We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,300
Open access books available

130,000
International authors and editors

155M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Linked Open Data: State-of-the-Art Mechanisms and Conceptual Framework

Kingsley Okoye

IntechOpen

Abstract

Today, one of the state-of-the-art technologies that have shown its importance towards data integration and analysis is the linked open data (LOD) systems or applications. LOD constitute of machine-readable resources or mechanisms that are useful in describing data properties. However, one of the issues with the existing systems or data models is the need for not just representing the derived information (data) in formats that can be easily understood by humans, but also creating systems that are able to process the information that they contain or support. Technically, the main mechanisms for developing the data or information processing systems are the aspects of aggregating or computing the metadata descriptions for the various process elements. This is due to the fact that there has been more than ever an increasing need for a more generalized and standard definition of data (or information) to create systems capable of providing understandable formats for the different data types and sources. To this effect, this chapter proposes a semantic-based linked open data framework (SBLODF) that integrates the different elements (entities) within information systems or models with semantics (metadata descriptions) to produce explicit and implicit information based on users' search or queries. In essence, this work introduces a machine-readable and machine-understandable system that proves to be useful for encoding knowledge about different process domains, as well as provides the discovered information (knowledge) at a more conceptual level.

IntechOpen

Keywords: LOD, semantics, ontologies, metadata creation, data integration, process description, information retrieval, information extraction, information systems

1. Introduction

Linked Open Data (LOD) is a term used to refer to tools or platforms that support freely-connected (interlinked) resources or frameworks to allow for collection and integration of data (usually derived from various sources or formats) and provide useful information that can be accessed by machines or humans. Typically, LOD supported tools or platforms is expected to allow for both simple or complex oriented lookup for information access through some form of predefined language or mechanisms (e.g. using scripts or query-based languages such SQL, HTML, SPARQL, Description Logics, RDF graphs of the Triples form, XML, etc.) [1–3]. Technically, LOD is semantically defined as a knowledge graph [3] that vents in the form of

semantical web or schema (e.g. using ontologies) [4–6] of interconnected data [7]. According to Snyder et al. [7], LOD has since been epitomized as a way of improving the process of discovering useful information or resources by creating a series of robust links between related concepts or items.

The work done in this chapter notes that one of the main challenges with LOD has been on how to create systems or methods that are capable of providing an understandable format (both machine-readable and machine-understandable) for the various datasets that may come from different sources, as well as, making the derived formats or standards explicable across the several platforms. To this end, the work proposes a semantic-based LOD framework (SBLODF) that provides an additional function to LOD that allows for formal integration of the process elements or concepts through metadata creation (process description) using the semantic technologies or schema. This is called Semantic-based Linked Open Data.

2. Preliminaries

2.1 Semantics? the missing link in LOD systems

Research on why "domain knowledge" is useful in bridging the semantic gap in existing systems or applications that aims to store and/or process data has long been discussed in existing works of literature [1, 6, 8–11]. Whereas, Declerck et al. [1] note that one of the main aims of LOD supported systems is to develop new ways or methods for construing data values (interlinks) that are applicable to a broad range of applications or platforms (based on language technologies or resource descriptions) through semantic technologies. Wang [12] notes that contemporary studies on LOD methods and tools are mainly directed towards ascertaining different levels or types of process instances (entities), thereby resulting in the central task of finding relationships (schema-level) or links that exist amongst the LOD datasets or models in question being ignored.

According to Wang [12], ontological representations (mappings) are a very crucial way of solving the data heterogeneity or missing link. Moreover, ontologies can be described as an essential tool that proves useful towards establishing the semantic-level links in LOD [6, 13–18]. For example, Selvan et al. [19] proposed an ontology-based recommender system that is built on cloud services to store and retrieve data for further analysis using Type-2 fuzzy logic.

Studies have shown that there exists a (semantic) gap between different datasets and the various tools/algorithms that are applied to analyze or understand the data including results of the analysis in all stages of the data processing; ranging from the data pre-processing to implementation of the algorithms, and the interpretation of the results [6, 8, 11, 16]. For instance, data pre-processing usually involves the process of filtering and cleaning of data, standardization by defining formats for its integration, transformation and properties extraction and retrieval of the defined formats/structures, and then selected for the purpose of analysis. Nevertheless, in many settings, there exist the issue of semantic gaps in the several phases of the data pre-processing. For example, we note that in the absence of considering the formal structure (semantics) of the data models, most of the resulting systems have resort to empirical or ad-hoc methods to determine the quality of the underlying datasets or concepts. Whereas, it is certain that data semantics is necessary for understanding the relations that exist amongst the different process elements in the models, especially during the standardization and transformation step. Thus far, it is important to determine the correlation between the different data elements by taking into account the underlying properties/attributes of the data when performing

data standardization or processing at large. Apparently, tightly (closely) correlated attributes can be generalized into one combined attribute or classification for the purpose of tractability and conceptualized analysis.

Typically, in terms of the different application domains and rule-based information extraction systems, Yankova [20] conducted a semantic-based identity resolution and experiment that aims to identify conceptual information expressed within a domain ontology. The experiment was based on a generic and adaptable human language technology. In the experimentation, they extracted company information from several sources and update the existing ontologies with the resolved entities. The method for information extraction is a rule-based system they referred to as Identity Resolution Framework (IdRF) built using Proton [20] that provides a general solution to identifying known and new facts in a certain domain, and can also be applied to other domains regardless of the type of entities that may need to be resolved. Moreover, input to the IdRF includes different entities together with their associated properties and values, and the expected output is an integrated representation of the entities that are consequently resolved to have new properties or values within the ontology.

On the one hand, ontologies have shown to be beneficial in such data processing or conceptualization scenarios [21–23]. Ontologies are formal structures that are used to capture knowledge about some specific domain processes of interest [24–25]. Technically, the "ontologies" or formal expressions (taxonomies) per se are used to describe concepts within process domains as well as the relationships that hold between those concepts. Ontologies range from the tools or mechanisms used to create the taxonomies, to the population of the classified elements or database schemas to fully axiomatized theories [11]. Practically, ontologies are used by the domain experts to (manually, semi-automatic, or automatically) fill the semantic gaps that are allied to the data analysis procedures and models.

On the other hand, it is also noteworthy to mention that ontologies are now central to many applications; such as scientific knowledge portals, information management and integration systems, electronic commerce and web services, etc. which are all grounded or built on the LOD scheme.

2.2 State-of-the-art: semantic schema for data integration and processing

Indeed, several areas of application and definition of ontologies (schema) have been noted in the current works of literature especially as it concerns the varied domains of interest. For example, Hashim [26] notes that the term "ontology" is borrowed from the philosophy field that is concerned with being or existence, and further mention that in context of computer and information science, it symbolizes as an "artefact that is designed to model any domain knowledge of interest". Ontology has also been broadly used in many sub-fields of the computer science and AI, particularly in data pre-processing, management, and LOD related areas such as intelligent information integration and analysis [27], cooperative information management systems [28], knowledge engineering and representation [29], information retrieval [30], information extraction [31], ontology-based information extraction systems [13, 15, 32–34], database management systems [35–37], and semantic-based process mining and analysis [10, 16, 18, 38–40].

Gruber [25] describes the ontological concept or notion as "a formal explicit specification of a conceptualization". To date, the aforementioned breadth has been the most widely applied and cited definition of ontologies within the computer science field. The description means that ontologies are able to explicitly define (i.e. specify) concepts and relationships that are paramount for modeling any given process or domain of interest. Moreover, with such expressive application

or schema, it means that the processes can be represented in the form of classes, relations, individuals, and axioms (C,R,I,A). Thus, we note that the structural layer of ontologies can be defined as a *quadruple* which are construed on connected sets of taxonomies (RDF + Axioms) or yet formal structure (Triple + Facts). Whereby the *subjects* include the represented class(es), *C*, the *objects* include the individual process elements or instances, *I*, the *predicates* are used to express the relationships, *R*, that exist amongst the subjects and objects, and then sets of axioms that state facts, *A*, [11]. Thus;

$$Ont = (C,R,I,A) \quad (1)$$

Following the aforementioned definition of the ontological concept or schema, this work note that ontologies serve and are built to perform the main functional mechanisms for the integration of data models for the various systems (e.g. LOD) as follows:

- *Conceptualisation:* method used to represent abstract models of a phenomenon in real-world settings. This is done by identifying suitable domain (semantic) relationships that exist amidst the process elements (concepts) through formal definitions in what can be called declarative axioms that allow for the resultant models to be represented (conceptualisation) declaratively.

- *Explicitness:* procedures that allow or support the different types of concepts and restrictions on their use (properties assertions) to be defined explicitly.

- *Formality:* expressions which are defined to prevent unexpected interpretation of the C,R,I,A as quadruple (e.g. concepts and notations, relationships, properties restrictions, etc.). Thus, it enables the resultant systems or models to be machine-readable and machine-understandable, respectively.

3. Proposed semantic-based LOD framework (SBLODF)

The representation (modeling) of knowledge using ontologies (e.g. taxonomies) helps in organizing *metadata* for complex information or data structures. According to Sheth et al. [41], description of real-time processes through metadata creation provides a syntactic as well as semantic way of representing information about the resources that are encoded as instances (entities) in ontological form. Besides, the formal representation of ontologies and the underlying metadata created as a result of the representations allows for automatic reasoning of the processes by making references (inference) to the defined concepts [42]. Indeed, with such reasoning aptitude, the process analysts or owners are able to ensure specification of the process domains (knowledge) in view in an ontological form that can logically be interpreted in an apt way. Consequently, this permits for automatic reasoning of the different concepts to derive an explicit/implicit knowledge about the process domains in question [43].

Therefore, the main benefits of ontologies for formal integration of datasets and models in any shape or platform can be summarized in two forms: (i) encoding knowledge about the specific process domains, and (ii) conceptual analysis and reasoning of the processes at more abstraction levels as described in detail in the following section.

3.1 Architecture of the SBLODF framework

Information retrieval and structuring of the different sets of data that are stored in several databases or knowledge-base are usually performed in alignment with the users' query [38]. As gathered in **Figure 1**, the supported formats may be a list of document files or keywords issued to the system through the query module (functional operators). In turn, the retrieval module references the properties descriptions (conceptual assertions) that underlie the (semantic) models to produce information that is relevant to the users' query. For example, using the superClass-subClass hierarchies that are usually defined in a taxonomical form in ontologies. This is done through the classification process (e.g. classifying by using a reasoner) to compute the relevant information (e.g. individual entities or process instances) that fulfills the properties restriction by definition [44]. Technically, the most fitting (related) concepts are then presented to the user in a formal way, e.g. explicitly and implicitly.

Furthermore, we note that information retrieval and extraction systems such as the SBLODF framework (**Figure 1**) typically do not only support unstructured data or documents (e.g. textual data), but it also deals with semi-structured and structured data. This is where the semantic technologies and such type of systems (which combines the information retrieval (IR) with information extraction (IE) features) [38] becomes greatly beneficial. Functionally, the resulting system allows for merging and manipulation of structured, semi-structured, and unstructured data through the search (query) modules by enabling a conceptual intersection or reasoning between the different elements as contained in the system. Thus, the SBLODF is referred to as a conceptualization method or information processing system that combines the features of the machine-readable and machine-understandable systems or mechanisms.

For example, enterprise vendors such as FAST (a Microsoft subsidiary) incorporated analytical search functions to support data visualization and reporting into

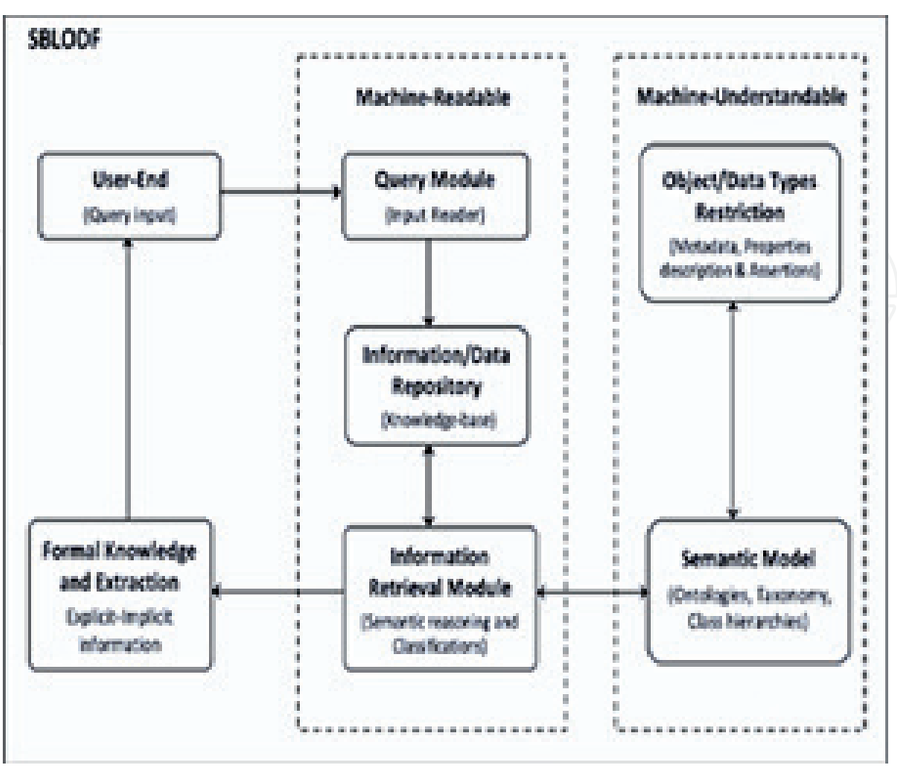

Figure 1.
Semantic-based linked open data framework (SBLODF).

their products [38, 45]. Moreover, Ingvaldsen [38] notes that the business process intelligence (BPI) solutions and offerings can also benefit from such a combination (IR and IE supported systems) by giving the users a search facilitated (data analysis) environment to harvest/harness data from both structured and unstructured data sources. Thus far, giving the users a more flexible environment for accessing relevant data items.

Interestingly, semantic-based information retrieval and extraction systems as illustrated in **Figure 1**, represents to be a step further in supporting the BPI's by providing additional modules or components that allows for integrating metadata description (e.g. ontologies) to the system design or framework. The semantic-based components (see: **Figure 1**) aims to add a machine tractable and/or re-purposeable layer of annotations that are relative to ontologies in order to complement the existing web of information and data analysis procedures, or yet, the omnipresence of natural language hypertext [4, 46, 47]. Perhaps, this is fundamentally done through the creation of semantic annotations [11, 23] and linking of the different concepts or modules to ontologies. In turn, the semantically motivated process or models turns out to become automatic or semi-automatic in nature and allows for ample integration of the LOD frameworks due to creation, interrelation, or application of the ontologies (semantic schema). Besides, this has led to the advancement of hybrid intelligent systems such as the ontology-based information extraction systems (OBIE) [9, 13, 15]. Explaining why IE and semantic technologies can be used to bring together a common language or syntax upon which the LOD systems or web search are built specially given the ever-needed formal knowledge or tools for information (data) access and utilization.

Some examples of state-of-the-art tools or systems that trails to support the semantic-based LOD framework or search include; KIM (knowledge and information management system) [31, 48] an extendable platform for information management that seemingly offers IE-based functions for metadata creation and search. Technically, KIM consists of a set of front-end (user-interface) for online information search by offering semantically-enhanced browsing features.

Another tool that tends to support the semantic-based LOD, such as the SBLODF framework described in this chapter, is Magpie [49]. Magpie is developed and implemented as an add-on to web browsers by using IE mechanisms to support collaborative information interpretation and modeling of the extracted knowledge from the web. As illustrated in **Figure 1**, it annotates the different web pages with metadata descriptions in an automated manner by automatically populating ontologies from the relevant (web) sources. Thus, the application (Magpie) is interoperable with ontologies or semantic schema. Moreover, it is important to mention that one of the fundamental elements of the tool (Magpie) that is pertinent to this work is the fact that it makes use of ontologies to provide specific (tailored content) information to the users.

There are several other platforms that can be referred to also support the SBLODF framework. This includes the SemTag [50] which utilizes IE facilities or function to support large scale semantic annotations and process descriptions using TAP ontology. As described in **Figure 1**, SemTag functions by performing annotation of all defined mentions (references) of any given process instance or entity in the ontology (TAP) through a lookup phase. This lookup process is then followed by the disambiguation phase during which it assigns the right classes (or establishes instances that do not correspond to a class in the TAP) using a vector-space model [50].

4. Implementation components of the semantic-based linked open data framework (SBLODF)

The work describes in this section (**Figure 2**) how the semantic schema is used to support the development of the LOD framework. Ontology-based information retrieval and extraction systems such as the SBLODF (**Figure 1**) are construed on the main building blocks [31]:

- *Named Entity recognition* (NE) which trails to find and classifies the different concepts that can be found within the model or knowledge-base.
- *Co-reference resolution* (CO) which identifies the relations or association that co-exist amongst the concepts or entities.
- *Template Element construction* (TE) that adds descriptive information (meta-data) to the classified NE through the CO component.
- *Template Relation construction* (TR) that locates the links or references between the TE (entities), and
- *Scenario Template production* (ST) that matches (fits) the TE and TR components into a specified scenario or process instance.

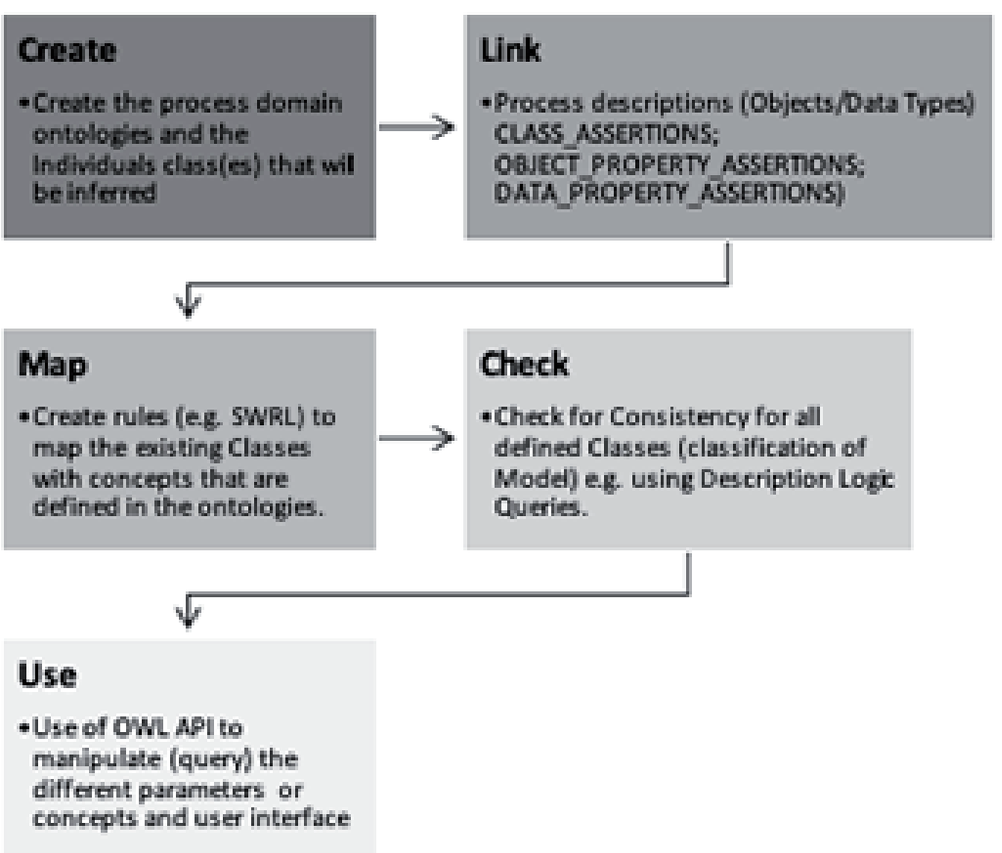

Figure 2.
Implementing the semantics components in SBLODF using create-link-map-check-use (CLMCU) procedure [11].

Interestingly, Dou et al. [8] note that a well-designed information retrieval or data/process mining system should present the outcomes or discovered information in a formal and structured format qua being interpreted as domain knowledge, or yet, utilized to further augment the existing system. Besides, the work [8] states that ontological schema is one of the most effective ways to formally represent any given type of data or process models. This is due to the fact that concepts defined within ontologies can be expressed or represented as set(s) of annotated terms and/or relations that aims to support information extraction and association rule mining systems especially with those allied to the ontology-based information and extraction (OBIE) [9].

To this effect, this current study note that to implement the aforementioned functionalities of the ontology-based systems in the SBLODF framework, the extracted information or models from the standard process mining (management) or analysis tools/sources needs to be represented as sets of annotated terms (that links or connects the defined terms) in an ontological form using the create-link-map-check-use (CLMCU) incremental or semantic modeling procedure [6, 11].

As illustrated in **Figure 2**, the resultant class hierarchies or taxonomy (ontologies) tends to provide a way of formally representing the defined (annotated) terms or concepts in a structured format by ascertaining the relationships (association) that co-exist amongst the several entities within the process model. Henceforth, the process descriptions and assertions are realized by encoding the process model in the formal structure or taxonomy, thus far ontologies, for the information/knowledge extraction to follow. In the end, the system is integrated or manipulated with an inference engine (e.g. reasoner or classifier) that performs semantic reasoning by uncovering the different levels of the ontological classification and process elements to produce the (inferred) information (knowledge) based on the input queries or users search that displays to be closer to human understanding (machine-understandable).

5. Data analysis and implementation results

For the data analysis and implementation in this section of the chapter; the work uses dataset about a real-time business process provided by the IEEE CIS Task Force on Process Mining [51] to illustrate how the proposed method is capable of performing the information retrieval and extraction process by integrating the different components of the SBLODF framework, as described in **Figure 1**. Typically, this is done by enabling a conceptual intersection or reasoning between the different elements/components which are supported by the system. These functions ranges from the user input query or search module to the information retrieval module or input reader (machine-readable component), and then, the metadata descriptions/assertions, ontological modeling and class hierarchies (taxonomy) to the provision of formal knowledge (explicit and implicit information) that can be easily understood by humans in real-world settings. Fundamentally, the work note the key function of the SBLODF framework to be in its capability to utilize the semantic concepts to perform automatic (semantic) reasoning/inferences capable of discovering useful models and conceptual information from the dataset. Henceforth, the SBLODF implementation allows the meaning of the process elements to be enhanced through the use of property description languages and classification of the discoverable entities, for example, using the Web Ontology Language (OWL) [4], Semantic Web Rule Language (SWRL) [52], and Description Logic (DL) [2].

Practically, as shown earlier in **Figure 2**, the ontological schema or framework trails to connect the different sets of discoverable entities in the model with their

class membership, or yet with a fixed literal, and can also describe the sub assumption hierarchies (taxonomies) that exists between the various classes including the relationships that they share within the underlying model. Moreover, the different class(es) are consequently instantiated with the set of individuals, I, and can also contain the various set of axioms, A, which states facts. For instance, the true positive elements, i.e., what is true and fitting within the model, and true negatives, i.e., what is true and not fitting in the model.

To illustrate this, the work analyzes the data provided in Ref. [51], by making use of the object properties (see: **Figure 2**) to describe the different classes that can be found within the semantic model developed with Protégé Editor for the purpose of this work. As shown in **Figure 3**, it used the "hasTraceFitness" object property to describe the classes or entities in the test data log that has a "TrueTrace_Classification_(TP)" or "FalseTrace_Classification_(TN)".

Moreover, as defined in Section 2.2 and Section 4 (**Figure 2**), if we Let A, be the set of all process executions or actions that can be performed within the semantic model. A process action $a \in A$ is characterized by a set of input parameters $Ina \in P$ which is required for the execution of a, and a set of output parameters $Outa \subseteq P$ which is produced by a after the execution or search query. Thus, with such function, the extraction and automatic reasoning (e.g. classification) of the process parameters is enabled and/or supported by the model. Perhaps, the key purpose of implementing the framework is to match the questions one would like to answer about attributes/relationships the process instances share amongst themselves within the knowledge-base by linking to the concepts (inferred classes) described in the model.

As shown in **Table 1**, based on the features of the provided datasets [51], the work applies the cross-validation technique to analyze the training and test sets. The traces were computed and recorded according to the *reasoner* response, and the classifier (reasoner) was tested on the resulting individuals by assessing its performance with respect to the correctly classified traces. As an example, the following DL queries/syntax [2] represents as set of input parameters (search query) the work executed in order to output the set of traces that can be found within the

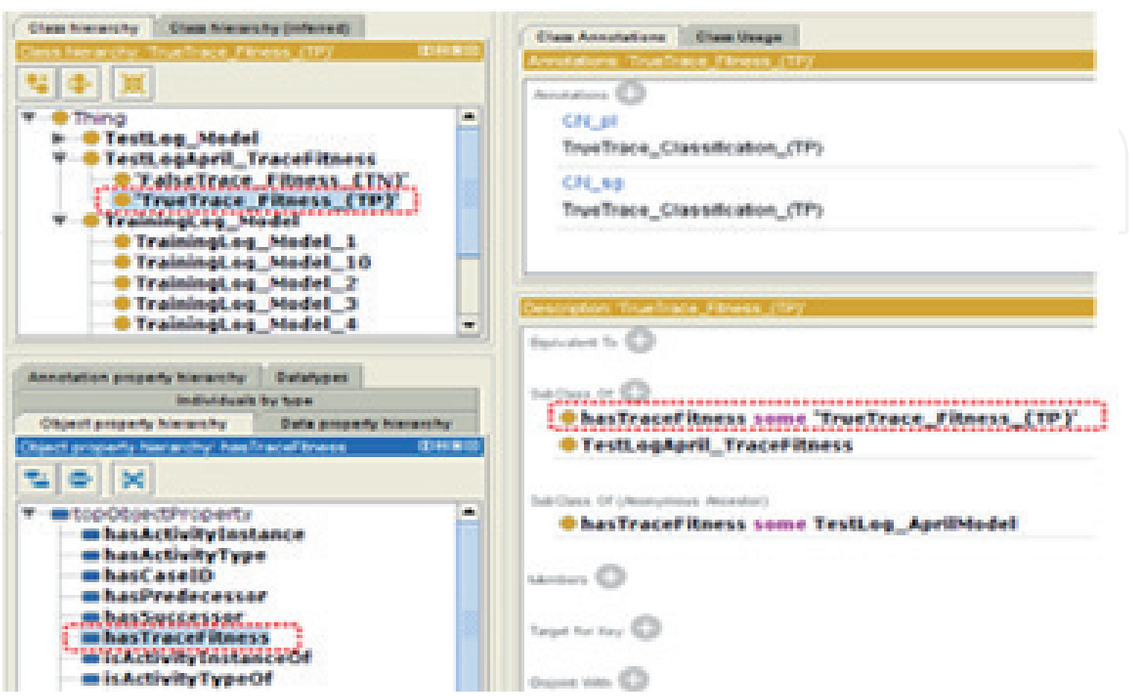

Figure 3.
Example of object property description and assertion for the true trace classification.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Trace_1	TP *	TN *	TP *	TN *	TN *	TN *	TP *	TP *	TP *	TP *
Trace_2	TN *	TN *	TP *	TP *	TP *	TP *	TP *	TN *	TP *	TP *
Trace_3	TP *	TP *	TP *	TN *	TN *	TN *	TN *	TP *	TP *	TN *
Trace_4	TP *	TP *	TN *	TP *	TN *	TP *	TN *	TP *	TP *	TN *
Trace_5	TN *	TN *	TN *	TP *	TN *	TP *	TN *	TP *	TP *	TN *
Trace_6	TP *	TN *	TN *	TP *	TN *	TP *	TP *	TN *	TN *	TP *
Trace_7	TN *	TP *	TP *	TN *	TN *	TP *	TN *	TP *	TN *	TN *
Trace_8	TN *	TP *	TP *	TP *	TN *	TN *	TP *	TP *	TP *	TP *
Trace_9	TP *	TN *	TP *	TN *	TP *	TN *	TP *	TP *	TN *	TP *
Trace_10	TP *	TN *	TP *	TN *	TN *	TN *	TP *	TP *	TP *	TP *
Trace_11	TN *	TP *	TP *	TP *	TP *	TN *	TN *	TN *	TN *	TP *
Trace_12	TP *	TN *	TN *	TP *	TP *	TP *	TP *	TN *	TP *	TN *
Trace_13	TP *	TP *	TN *	TN *	TP *	TN *	TN *	TN *	TN *	TP *
Trace_14	TN *	TP *	TN *	TN *	TN *	TN *	TN *	TP *	TN *	TP *
Trace_15	TP *	TN *	TN *	TN *	TP *	TP *	TN *	TN *	TN *	TN *
Trace_16	TN *	TN *	TN *	TP *	TP *	TN *	TN *	TN *	TP *	TN *
Trace_17	TP *	TP *	TP *	TP *	TP *	TP *	TP *	TN *	TN *	TP *
Trace_18	TN *	TP *	TN *	TN *	TP *	TP *	TP *	TN *	TN *	TN *
Trace_19	TN *	TP *	TP *	TP *	TN *	TP *	TP *	TP *	TN *	TN *
Trace_20	TN *	TN *	TN *	TN *	TP *	TN *	TN *	TN *	TP *	TN *
True positives (TP):	10	10	10	10	10	10	10	10	10	10
False positives (FP):	0	0	0	0	0	0	0	0	0	0

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
True negatives (TN):	10	10	10	10	10	10	10	10	10	10
False negatives (FN)	0	0	0	0	0	0	0	0	0	0
No. of traces correctly classified	20	20	20	20	20	20	20	20	20	20

Note: cells with gold sign (*) indicates traces that were correctly classified by the reasoner which equals to 200 traces out of 200.

Table 1.
Classification results and performance of the discovered models.

intechopen

defined model that has "TrueTrace_Fitness_(TP)" and "FalseTrace_Fitness_(TN)" respectively.

"TestLog_(forSpecifiedClass) and hasTraceFitness some
'TrueTrace_Fitness_(TP)".
"TestLog_(forSpecifiedClass) and hasTraceFitness some
'FalseTrace_Fitness_(TN)".

Thus, as reported in **Table 1**, each results of the classification process for the discovered models, i.e., the true positives and true negatives traces, were determined.

From the results of the classification method (**Table 1**), we note for each run set of parameters retrieved from the model that the commission error, otherwise referred to as error-rate (false positives (FP) and false negatives (FN)) was null, thus, equal to 0. This means that the reasoner (classifier) did not make critical mistakes. For instance, a case whereby a trace could be considered to be an instance of a class while it is categorically an instance of another class. In the same vein, the work notes that the accuracy rate (i.e., true positives (TP) and true negatives (TN)) when determining the different traces and classifications was very high, thus, correct, and were consistently observed for all the test sets.

6. Discussion and conclusion

LOD systems or frameworks and algorithms are fundamentally aimed to provide a standard platform for integrating/analyzing different datasets or models to extract snippets of information that are relevant to the users, independent of the various formats or syntax. In other words, LOD stands as the bridge between the different data formats/sources and knowledge acquisition or information retrieval. For example, Cunningham [31] notes that the process of extracting information from the several sources may simply imply taking text documents, speech, graphics, etc., as input and produces fixed-format unambiguous data (or information) as output. In turn, the discovered information or data may be directly displayed to the users, stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in IR-supported applications such as the web search technologies, internet, or search engines like Google, Bing, etc.

Studies have shown that IE technologies may be distinctive from the IR systems or functions. Whereas, Cunningham [31] notes that the IR systems aims to find relevant information (e.g. texts) and presents them to the users, an IE application analyses the texts and presents only the specific information from the text that the user is interested in. Apparently, this kind of tailored information analysis is where ontology-based information extraction systems such as the SBLODF framework described in this chapter construes its incentives.

For example, a user of an IR-supported system wanting information on higher educational institutions that offers a particular course would enter a list of relevant words or keywords in the search module and receive in return a set of documents (e.g. various university prospectus, course guidelines, etc.) that contain likely matches based on the keywords. In turn, the user would read through the matches or documents and extract the requisite information they need themselves, or yet store them on their computer storage for future reference. Nonetheless, unlike IR, an IE system would automatically populate a list of tables or spreadsheets directly with the names of relevant universities and their course offerings making it easier for the users to extract or learn the specific information they need or seek to acquire.

However, there may also exist some limitations with IE supported LOD frameworks or systems when compared to IR only. One of the limitations is that IE systems are more difficult and knowledge-intensive to build and are to a certain extent tied to particular domains or case scenarios. Also, IEs are more computationally intensive than IRs. Although, on the other hand, when compared to applications where there are large text or document volumes, IEs are potentially much more efficient than IRs due to the capacity of dramatically reducing the amount of time people may spend reading through text documents to find the relevant information. Perhaps, the aforementioned benefit of the IEs is only possible as a result of applying the ontological (semantics) schema to represent and manipulate the underlying information as described in Section 3 of this chapter.

Moreover, in settings where the results need to be presented, for example, in several languages; the fixed-format and unambiguous nature of IEs outputs make the information retrieval process relatively direct when compared to the full translation facilities that are consequently needed for interpretation of the multilingual texts found by IRs. Indeed, this means that IEs only present the specific information in a form that the user is interested in, and this feature is where the ontology-based IE systems are more powerful given that ontology is one of such tools that have the capability of providing information in a structured format. For instance, the automatic population of the different class hierarchies in ontologies within OBIE [9] applications is capable of formally identifying process instances or element within a text file that belongs to or references certain concepts in the pre-defined ontologies, and then trails to add those instances to the model in the right locations.

Having said that, we note that OBIE systems such as the SBLODF attempts to classify the several entities in a more scalar way; as there may be different categories to which an entity can belong to and cataloging the discrepancies between those classifications is more or less straightforward when using the OBIE framework [17].

Furthermore, to explain the application of the OBIE concept in the context of information retrieval and extraction or semantic-based knowledge representation, Yankova [20] refers to an identity resolution method of deciding whether an instance extracted from a text by an IE application refers to a known entity within a target domain ontology. Technically, the authors [20] developed a customizable rule-based framework for identity resolution and merging that uses ontologies for knowledge representation by using customizable identity criteria put in place to decide on the similarity between two process instances or entities. The criteria utilizes ontological operations and similarity computation between extracted and stored values that are weighted. Besides, the weighting criteria are routinely specified according to the type of entities and the application domain.

Accordingly, studies have also shown that aggregation of the extracted information from the different data sources has greater advantages (e.g. complementing partial information from one source to another, increasing the confidence of the extracted information, and storage of updated information within the knowledge bases) [11, 14, 15, 17, 20, 23, 53]. Truly, the resultant methods prove to provide standard structures for resolving the identities or properties description of the different class(es) of entities (process instances) by using ontologies as the core (fundamental) knowledge representation tools that help to provide the formal descriptions that are complemented with semantics.

Interestingly, Yankova [20] reveals that one fundamental problem to be addressed when providing a structure for distribution of the conceptual knowledge such as with OBIE systems; is that of identifying and merging the instances extracted from the multiple sources. Basically, the process should aim at identifying newly extracted facts, e.g. from the derived models, and linking them to

their previous references or mentions. To this effect, we note that ontology-based systems, in general, poses two main challenges that are directed towards [31]:

- identification of the concepts (e.g., process instances or entities) within the ontologies, and

- automatic population of the ontologies with newly (inferred or classified) instances.

Perhaps, it is also important to mention that when the ontologies are populated with the process instances or concepts assertions; the ultimate function of the resultant (OBIE) systems would simply be to manipulate the process elements, for example, by uncovering the relationships that exist amongst the process instances and revealing those to the users or search initiators based on the query modules [6, 9, 16, 31, 44, 54]. Moreover, for rule-based systems like OBIE, such procedures are relatively unswerving. But for learning-based IE systems, it appears to be more problematic due to the fact that training data are most often required to train the models, and collecting the necessary training data is, on the other hand, likely to be cumbersome/bottleneck [31]. Although to resolve such issues, new training datasets may need to be created either manually or semi-automatically; which are a lot of the time is time-consuming and/or burdensome task.

However, new and emerging systems/methods are being developed with the aim to help address such *metadata creation* problems for knowledge management or data analysis to support the IE and LOD at large [1, 11–15, 23, 33, 55, 56]. Moreover, unlike the traditional IE systems where the extracted facts (or information) are only classified as belonging to pre-defined types, an ontology-based (semantic) IE system (such as the SBLODF) seeks to identify, analyze and represent information at the conceptual (abstraction) levels by establishing a link (references) between the entities residing in the underlying systems' knowledge-bases and their mentions within the contextual domain. Henceforth, semantically-based LOD systems should not only support the formal representation of the different domains. But should also, on the other hand, provide information about the several known entities and their properties descriptions. Thus, ontology-based LOD systems such as the SBLODF introduced in this chapter must integrate well-defined entities with their semantic descriptions for an efficient explicit and implicit information extraction and/or analysis, i.e., machine-readable and machine-understandable system.

Acknowledgements

The author would like to acknowledge the technical support of Writing Lab, TecLabs, Tecnologico de Monterrey, in the publication of this work.

DOI: http://dx.doi.org/10.5772/intechopen.94504

IntechOpen

IntechOpen

Author details

Kingsley Okoye
Writing Lab, TecLabs, Office of the Vice President for Research and Technology Transfer, Tecnologico de Monterrey, Monterrey, CP 64849, Nuevo Leon, Mexico

*Address all correspondence to: kingsley.okoye@tec.mx

IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] T. Declerck et al., "Recent Developments for the Linguistic Linked Open Data Infrastructure," in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 5660-5667.

[2] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. The Description Logic Handbook: Theory, Implementation and Applications, 2nd Ed. Cambridge University Press; 2007

[3] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, and K. Srinivas, "SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems," Springer, Cham, 2020, pp. 514-530.

[4] S. Bechhofer et al., "OWL Web Ontology Language Reference," Technical report W3C Proposed Recommendation, Manchester, UK, 2004.

[5] C. D'Amato, N. Fanizzi, and F. Esposito, "Query answering and ontology population: An inductive approach," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5021 LNCS, pp. 288-302, doi: 10.1007/978-3-540-68234-9_23.

[6] K. Okoye, S. Islam, and U. Naeem, "Ontology: Core Process Mining and Querying Enabling Tool," in Ontology in Information Science, C. Thomas, Ed. IntechOpen, 2018, pp. 145-168.

[7] Snyder E, Lorenzo L, Mak L. Linked open data for subject discovery: Assessing the alignment between Library of Congress vocabularies and Wikidata. In: International Conference on Dublin Core and Metadata Applications. 2019

[8] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," in Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015, 2015, pp. 244-251, doi: 10.1109/ICOSC.2015.7050814.

[9] Wimalasuriya DC, Dou D. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science. Jun. 2010;**36**(3):306-323. DOI: 10.1177/0165551509360123

[10] A. K. A. De Medeiros and W. M. P. Van Der Aalst, "Process mining towards semantics," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4891 LNCS, T. Dillon, E. Chang, R. Meersman, and K. Sycara, Eds. Springer, Berlin, Heidelberg, 2009, pp. 35-80.

[11] Okoye K, Islam S, Naeem U, Sharif MS, Sharif MhD S. Semantic-based process mining technique for annotation and modelling of domain processes. Int. J. Innovative Computing & Information Control. 2020;**16**(3):899-921

[12] Wang T. Aligning the large-scale ontologies on schema-level for weaving Chinese linked open data. Cluster Comput. Mar. 2019;**22**(2):5099-5114. DOI: 10.1007/s10586-018-1732-z

[13] D. Calvanese, M. Montali, A. Syamsiyah, and W. M. P. van der Aalst, "Ontology-driven extraction of event logs from relational databases," in Lecture Notes in Business Information Processing, 2016, vol. 256, pp. 140-153, doi: 10.1007/978-3-319-42887-1_12.

[14] De Giacomo G, Lembo D, Lenzerini M, Poggi A, Rosati R. Using

ontologies for semantic data integration. In: Flesca S, Greco S, Masciari E, Saccà D, editors. A Comprehensive Guide through the Italian Database Research over the Last 25 Years. Springer: Cham; 2018. pp. 187-202

[15] D. Calvanese, T. E. Kalayci, M. Montali, and S. Tinella, "Ontology-based data access for extracting event logs from legacy data: The onprom tool and methodology," in Lecture Notes in Business Information Processing, vol. 288, W. Abramowicz, Ed. Springer Verlag, 2017, pp. 220-236.

[16] A. K. A. de Medeiros, W. van der Aalst, and C. Pedrinaci, "Semantic process mining tools: core building blocks," in ECIS, Ireland, June 2008, 2008, pp. 1953-1964.

[17] Maynard D, Peters W, Li Y. Evaluating evaluation metrics for ontology-based applications: Infinite reflection. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). 2008

[18] A. H. Cairns, J. A. Ondo, B. Gueni, M. Fhima, M. Schwarcfeld, C. Joubert and N. Khelifa, "Using semantic lifting for improving educational process models discovery and analysis," in CEUR Workshop Proceedings, 2014, pp. 150-161.

[19] Selvan NS, Vairavasundaram S, Ravi L. Fuzzy ontology-based personalized recommendation for internet of medical things with linked open data. Journal of Intelligent Fuzzy Systems. Jan. 2019;**36**(5):4065-4075. DOI: 10.3233/JIFS-169967

[20] Yankova M, Saggion H, Cunningham H. Semantic-Based Identity Resolution and Merging for Business Intelligence. UK: Sheffield; 2008

[21] N. Khasawneh and C. C. Chan, "Active user-based and ontology-based Web log data preprocessing for Web usage mining," in Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings), WI'06, 2006, pp. 325-328, doi: 10.1109/WI.2006.32.

[22] D. Perez-Rey, A. Anguita, and J. Crespo, "OntoDataClean: Ontology-based integration and preprocessing of distributed data," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006, vol. 4345 LNBI, pp. 262-272, doi: 10.1007/11946465_24.

[23] K. Okoye, "Technique for annotation of fuzzy models: A semantic fuzzy mining approach," in Frontiers in Artificial Intelligence and Applications, 2019, vol. 320, pp. 65-75, doi: 10.3233/FAIA190166.

[24] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition. Jun. 1993;**5**(2):199-220. DOI: 10.1006/knac.1993.1008

[25] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum. - Comput. Stud. Nov. 1995;**43**(5-6):907-928. DOI: 10.1006/ijhc.1995.1081

[26] Hashim H. Ontological structure representation in reusing ODL learning resources. Asian Assoc. Open Univ. J. Aug. 2016;**11**(1):2-12. DOI: 10.1108/aaouj-06-2016-0008

[27] Seng JL, Kong IL. A schema and ontology-aided intelligent information integration. Expert Systems with Applications. Sep. 2009;**36**(7):10538-10550. DOI: 10.1016/j.eswa.2009.02.067

[28] Ouksel AM, Sheth A. Semantic interoperability in global information systems: A brief introduction to the research area and the special section.

SIGMOD Rec. Dec. 1999;**28**(1):5-12. DOI: 10.1145/309844.309849

[29] Brewster C, O'Hara K. Knowledge representation with ontologies: Present challenges-future possibilities. International Journal of Human Computer Studies. Jul. 2007;**65**(7):563-568. DOI: 10.1016/j.ijhcs.2007.04.003

[30] Manning CD, Raghavan P, Schutze H. Introduction to Information Retrieval. Cambridge University Press; 2008

[31] Cunningham H. Information Extraction, Automatic. UK: Sheffield; 2005

[32] H. M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature," PLoS Biol., vol. 2, no. 11, Nov. 2004, doi: 10.1371/journal.pbio.0020309.

[33] H. M. Müller, K. M. Van Auken, Y. Li, and P. W. Sternberg, "Textpresso Central: A customizable platform for searching, text mining, viewing, and curating biomedical literature," BMC Bioinformatics, vol. 19, no. 1, Mar. 2018, doi: 10.1186/s12859-018-2103-8.

[34] S. A. Hosseini, A.-R. H. Tawil, H. Jahankhani, and M. Arandi, "Towards an Ontological Learners' Modelling Approach for Personalised E-Learning," Int. J. Emerg. Technol. Learn., vol. 8, no. 2, p. 4, 2013.

[35] Alkharouf NW, Jamison DC, Matthews BF. Online analytical processing (OLAP): A fast and effective data mining tool for gene expression databases. Journal of Biomedicine & Biotechnology. Jun. 2005;**2005**(2):181-188. DOI: 10.1155/JBB.2005.181

[36] *A. Poggi*, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati, "Linking data to ontologies," in Journal on Data Semantics X., vol. 4900 LNCS, S. Spaccapietra, Ed. Springer Verlag, 2008, pp. 133-173.

[37] C. Snae and M. Brückner, "Ontology-Driven E-Learning System Based on Roles and Activities for Thai Learning Environment," Interdiscip. J. e-Skills Lifelong Learn., vol. 3, pp. 001-017, 2007, doi: 10.28945/382.

[38] Ingvaldsen JE. Semantic Process Mining of Enterprise Transaction Data. Norway; 2011

[39] K. Okoye, A. R. H. Tawil, U. Naeem, S. Islam, and E. Lamine, "Using semantic-based approach to manage perspectives of process mining: Application on improving learning process domain data," in 2016 IEEE International Conference on Big Data, BigData2016, 2016, Washington DC, USA, pp. 3529-3538, doi: 10.1109/BigData.2016.7841016.

[40] Okoye K, Naeem U, Islam S. Semantic fuzzy mining: Enhancement of process models and event logs analysis from syntactic to conceptual level. Int. J. Hybrid Intell. Syst. Nov. 2017;**14**(1-2):67-98. DOI: 10.3233/his-170243

[41] Sheth A, Bertram C, Avant D, Hammond B, Kochut K, Warke Y. Managing semantic content for the web. IEEE Internet Computing. Jul. 2002;**6**(4):80-87. DOI: 10.1109/MIC.2002.1020330

[42] P. Dolog and W. Nejdl, "Semantic web technologies for the adaptive web," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2007, vol. 4321 LNCS, pp. 697-719, doi: 10.1007/978-3-540-72079-9_23.

[43] Yarandi M. Semantic Rule-Based Approach for Supporting Personalised

Adaptive E-Learning. United Kingdom: University of East London; 2013

[44] K. Okoye, A. R. H. A.-R. H. Tawil, U. Naeem, and E. Lamine, "Discovery and enhancement of learning model analysis through semantic process mining," Int. J. Comput. Inf. Syst. Ind. Manag. Appl., 2016, vol. 8(2016), pp. 093-114

[45] De Leoni M, Adams M, Van Der Aalst WMP, Ter Hofstede AHM. Visual support for work assignment in process-aware information systems: Framework formalisation and implementation. Decision Support Systems. Dec. 2012;**54**(1):345-361. DOI: 10.1016/j.dss.2012.05.042

[46] Fensel D, Hendler JA, Lieberman H, Wahlster W, Berners-Lee T, Lieberman H. Spinning the Semantic Web : Bringing the World Wide Web to its Full Potential. MIT Press; 2003

[47] J. Davies, D. Fensel, and F. Van Harmelen, Towards the semantic web : ontology-driven knowledge management. J. Wiley, 2003.

[48] Popov B, Kiryakov A, Ognyanoff D, Manov D, Kirilov A. KIM - a semantic platform for information extraction and retrieval. Natural Language Engineering. Sep. 2004;**10**(3-4):375-392. DOI: 10.1017/S135132490400347X

[49] J. Domingue, M. Dzbor, and E. Motta, "Magpie: supporting browsing and navigation on the semantic web," in Proceedings of the 9th international conference on Intelligent user interface - IUI '04, 2004, pp. 191-197, doi: 10.1145/964442.964479.

[50] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. V. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, J. Y. Zien, "SemTag and seeker: Bootstrapping the semantic web via automated semantic annotation," in Proceedings of the 12th International Conference on World Wide Web, WWW 2003, 2003, pp. 178-186, doi: 10.1145/775152.775178.

[51] J. Carmona, M. de Leoni, B. Depair, and T. Jouck, "IEEE CIS Task Force on Process Mining - Process Discovery Contest", Ist Edition, 2016 https://www.win.tue.nl/ieeetfpm/doku.php?id=shared:edition_2016

[52] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, "SWRL: A Semantic Web Rule Language Combining OWL and RuleML", W3C Member Submission. https://www.w3.org/Submission/SWRL/

[53] K. Okoye, S. Islam, U. Naeem, M. S. M. S. Sharif, M. A. M. A. Azam, and A. Karami, "The application of a semantic-based process mining framework on a learning process domain," in Advances in Intelligent Systems & Computing, 2019, vol. 868, pp. 1381-1403, doi: 10.1007/978-3-030-01054-6_96.

[54] Okoye K, Tawil ARH, Naeem U, Lamine E. A semantic reasoning method towards ontological model for automated learning analysis. Advances in Intelligent Systems & Computing. 2016;**419**:49-60

[55] Okoye K, Applications and Developments in Semantic Process Mining. IGI Global Publishers. Hershey. USA. 2020

[56] Polyvyanyy A, Ouyang C, Barros A, van der Aalst WMP. Process querying: Enabling business intelligence through query-based process analytics. Decision Support Systems. Aug. 2017;**100**:41-56. DOI: 10.1016/j.dss.2017.04.011